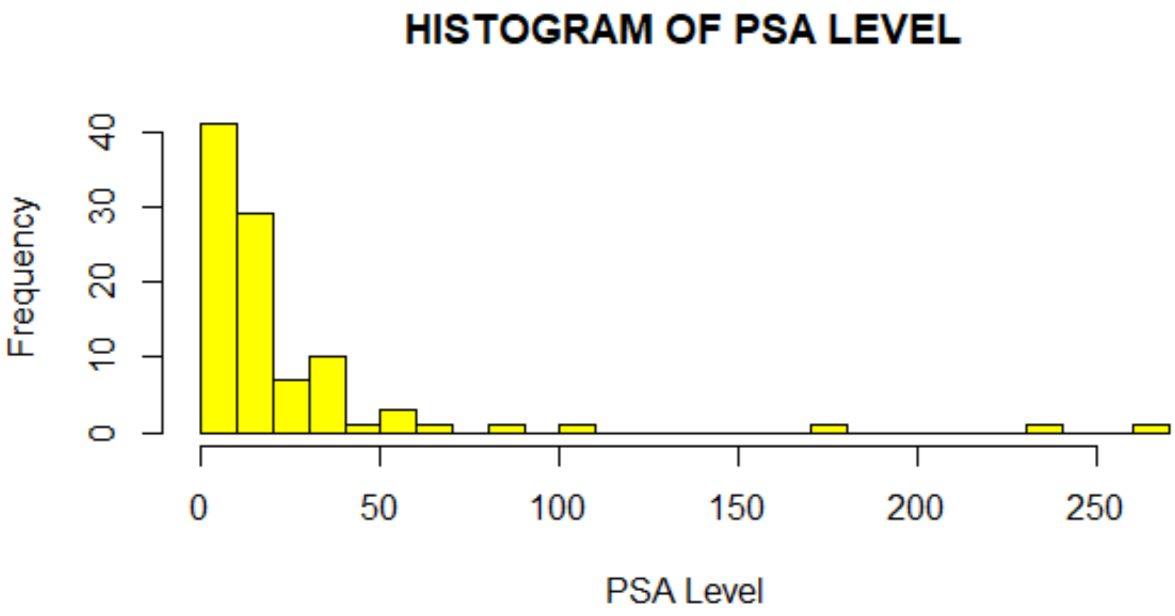
# Statistical Methods for Data Science: Mini Project 6 Solved

**Mini Project #:** 6
**Group #:** 9
**Names of group members:** Nikita Ahuja, Bhargaw Rajnikant Patel
**Contribution of each group member:** Both team members worked together to solve the problem. Both members worked out the solutions in R, wrote the code and finished the report in a timely manner. Both partners worked equally to complete the Mini Project 6 requirements.

**Problem 1:** Consider the prostate cancer dataset available on eLearning as prostate cancer.csv. It consists of data on 97 men with advanced prostate cancer. A description of the variables is given in Figure 1. We would like to understand how PSA level is related to the other predictors in the dataset. Note that vesinv is a qualitative variable. You can treat gleason as a quantitative variable.
Build a reasonably good linear model for these data by taking PSA level as the response variable. Carefully justify all the choices you make in building the model. Be sure to verify the model assumptions. In case a transformation of response is necessary, try the natural log transformation. Use the _nal model to predict the PSA level for a patient whose quantitative predictors are at the sample means of the variables and qualitative predictors are at the most frequent category.

Solution:
We first perform an exploratory analysis of the response variable – PSA level.

1. Histogram of PSA level:



From the above histogram we can conclude the following results:

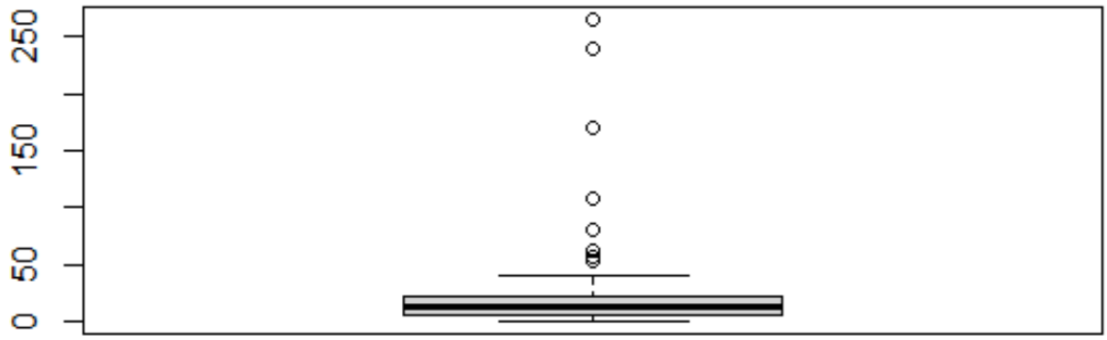
1. Many people have very Low PSA levels.

2. There is an indirect relation between the PSA levels and the number of people, because as the PSA level increases there is a drastic reduction in the number of people.

3. The distribution appears to be exponential and very different from a normal distribution.

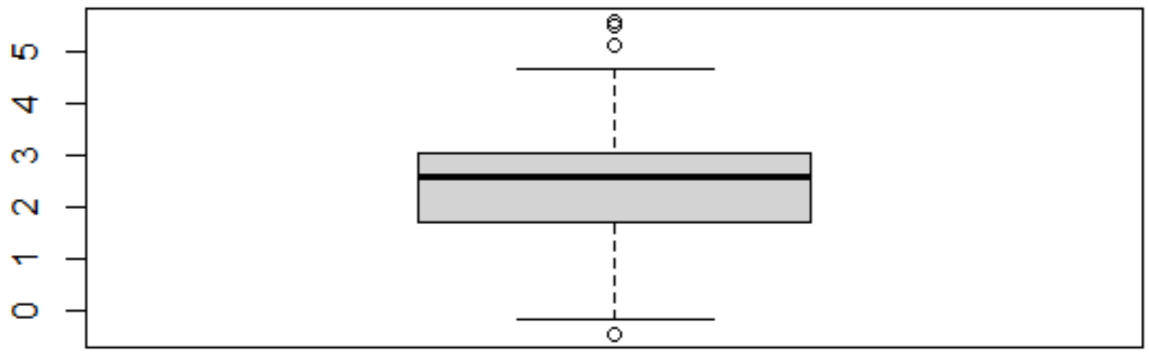
4. There is no strong evidence of the presence of outliers in the PSA data, so we plot the Boxplot.

2. Boxplot of PSA:



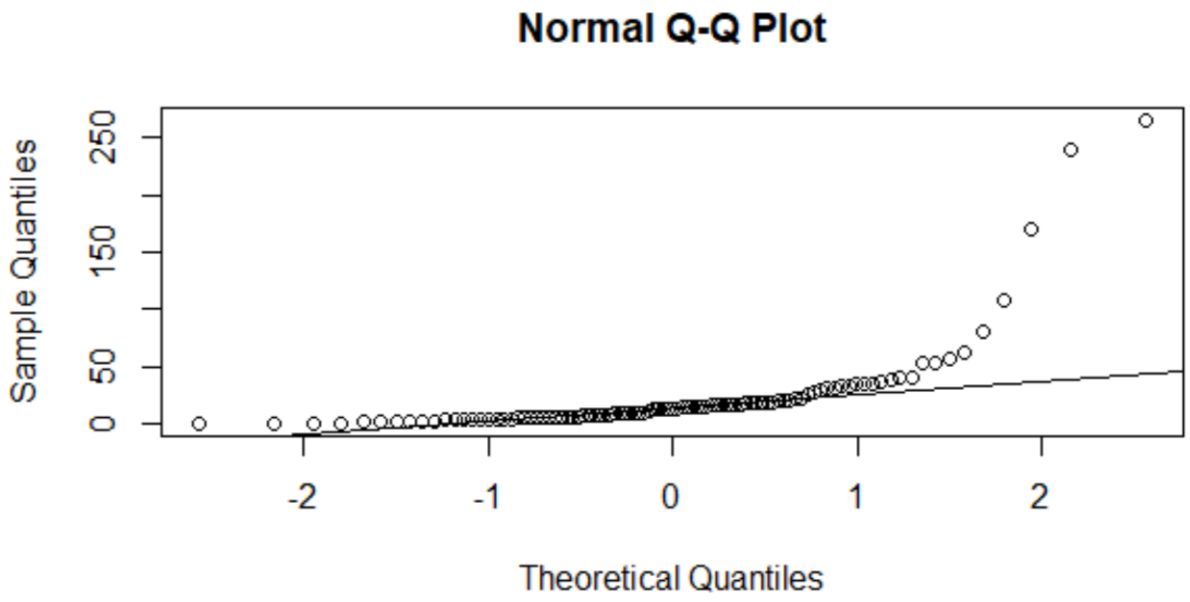

From the above boxplot, we can observe the presence of many outliers in the PSA data and we don't observe any symmetry here. For more detailed observation, let us look at the boxplot of the natural log transformation of the response variable.

3. Boxplot of Natural Log Transformed PSA:




From the above boxplot we observe that the distribution has now become symmetric and the number of outliers present in the PSA data have reduced. As the distribution is more symmetric here and the presence of outliers is less than what we observed from the normal boxplot, we now use the transformed response as our response variable.

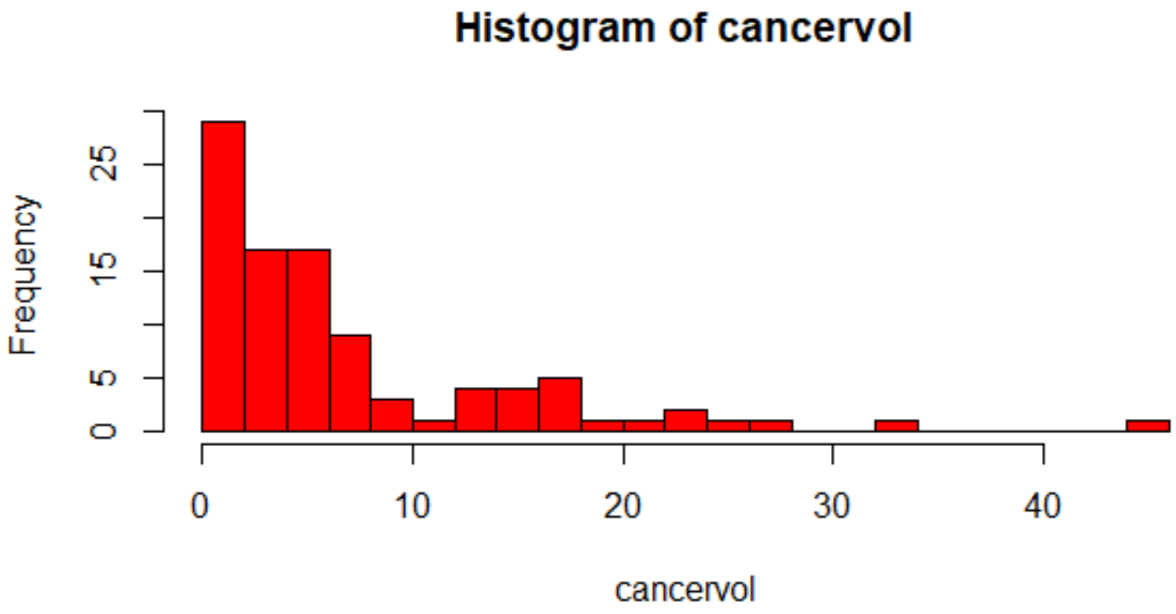
4. Normal Q-Q Plot of PSA Level:




From the above normal Q-Q plot of psa level we conclude that the data deviates from the normal Q-Q line and hence the given data does not approximate to the normal distribution.

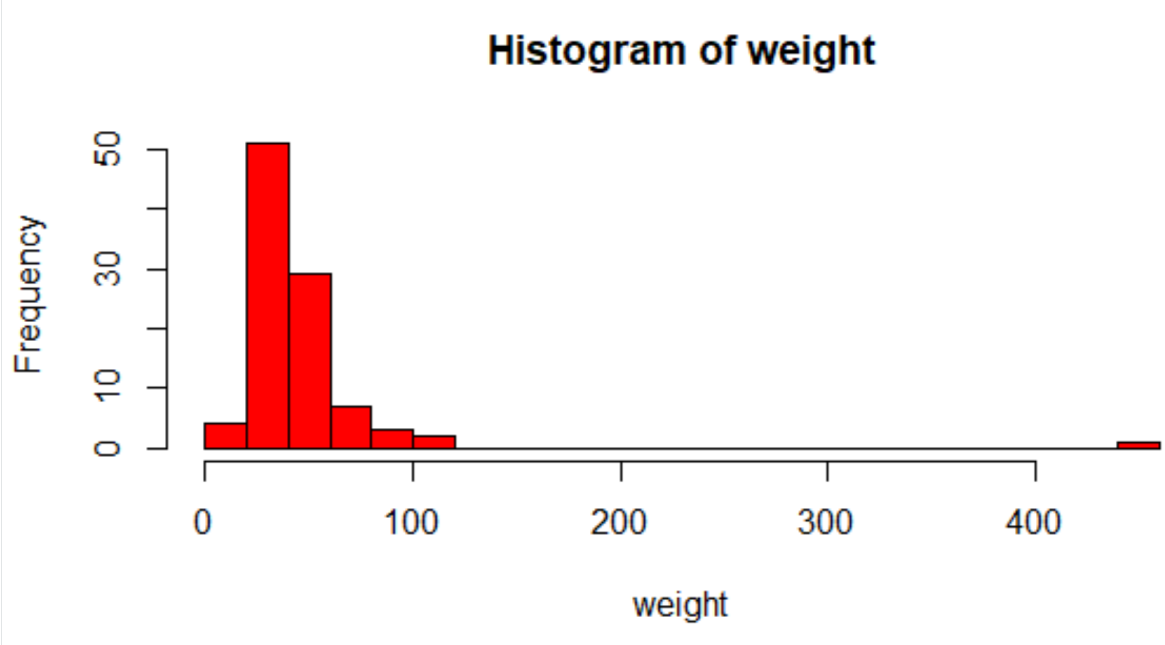
We now perform an exploratory analysis of the quantitative predictor variables.

1. Histogram of cancervol:



From the above histogram we observe that the histogram of cancervol is very similar to the histogram of PSA level as there distribution appears to be same i.e. Exponential Distribution. Hence, we can say that there might be a very strong linear relationship between the two.

2. Histogram of weight:

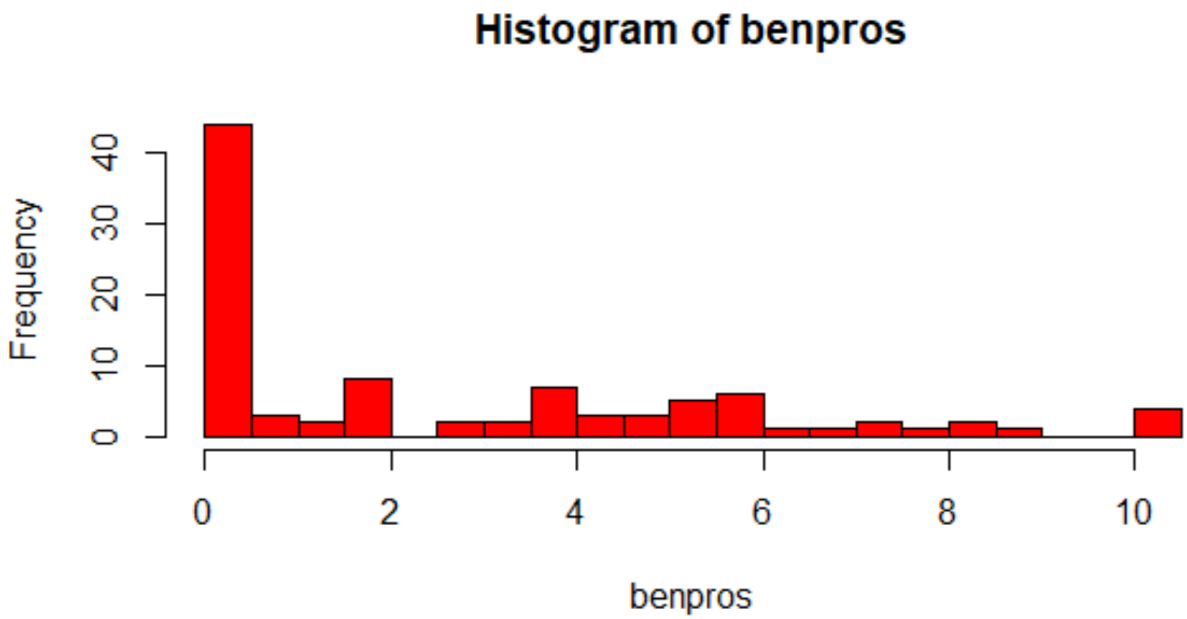



The above histogram is not an approximation to normal distribution but appears to be a gamma distribution.

3. Histogram of age:

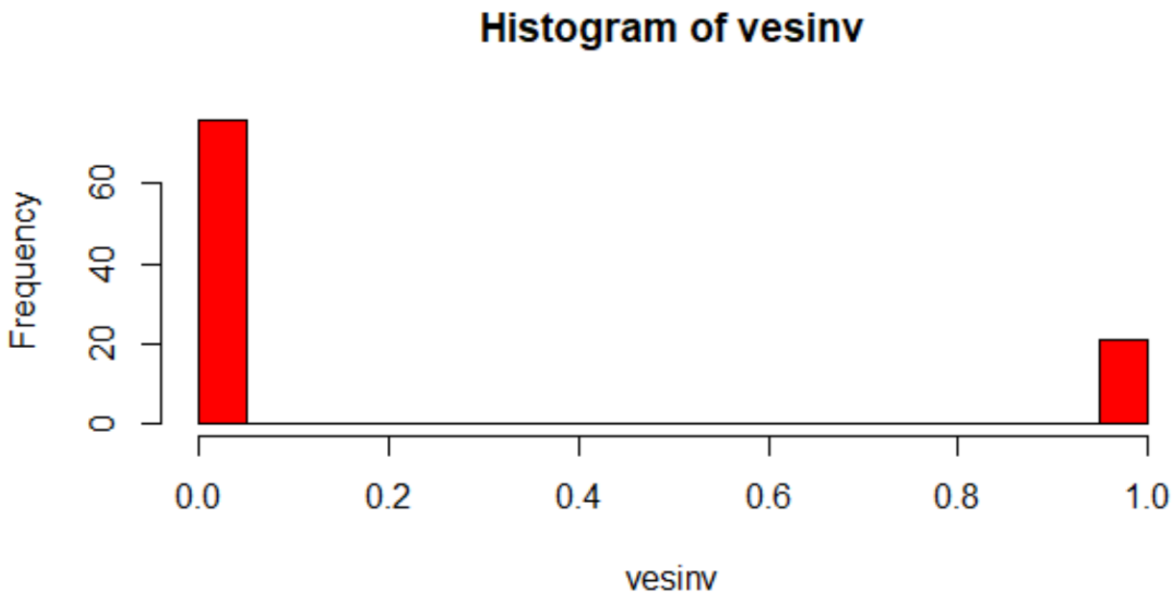

The above histogram of age approximates to a normal distribution which is expected for a random sample from a human population.

4. Histogram of benpros:



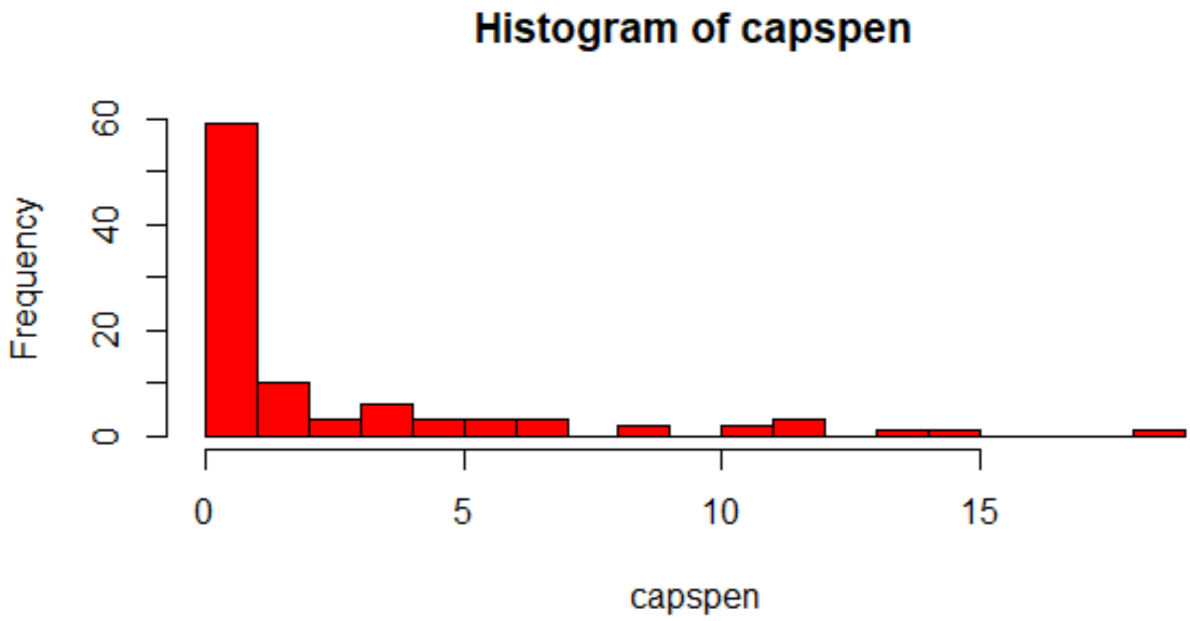

The above histogram of benpros is very similar to the histogram of PSA level and cancervol, as all of these histograms appears to be an exponential distribution. Here, we can say that there might exist a strong linear relationship between PSA level and benpros.

5. Histogram of vesinv:



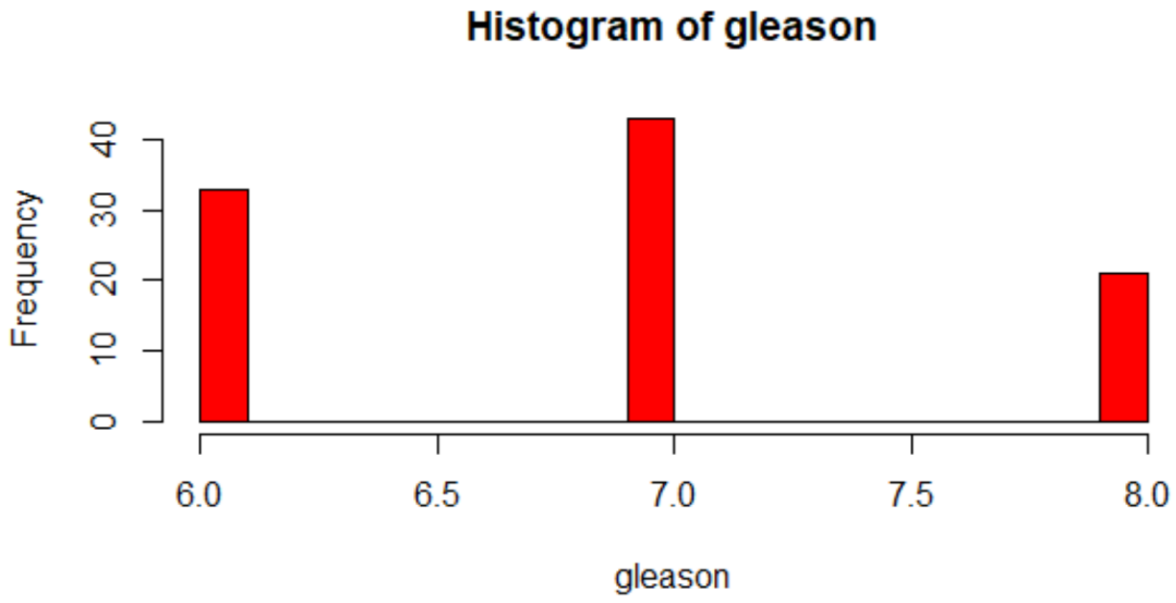

As vesinv is a qualitative variable, there exists only two possible values i.e. 0 or 1. From the above histogram we observe that there are very few people with Seminal Vesicle Invasion i.e. with value 1 than the people without Seminal Vesicle Invasion i.e. with value 0.

6. Histogram of capspen:



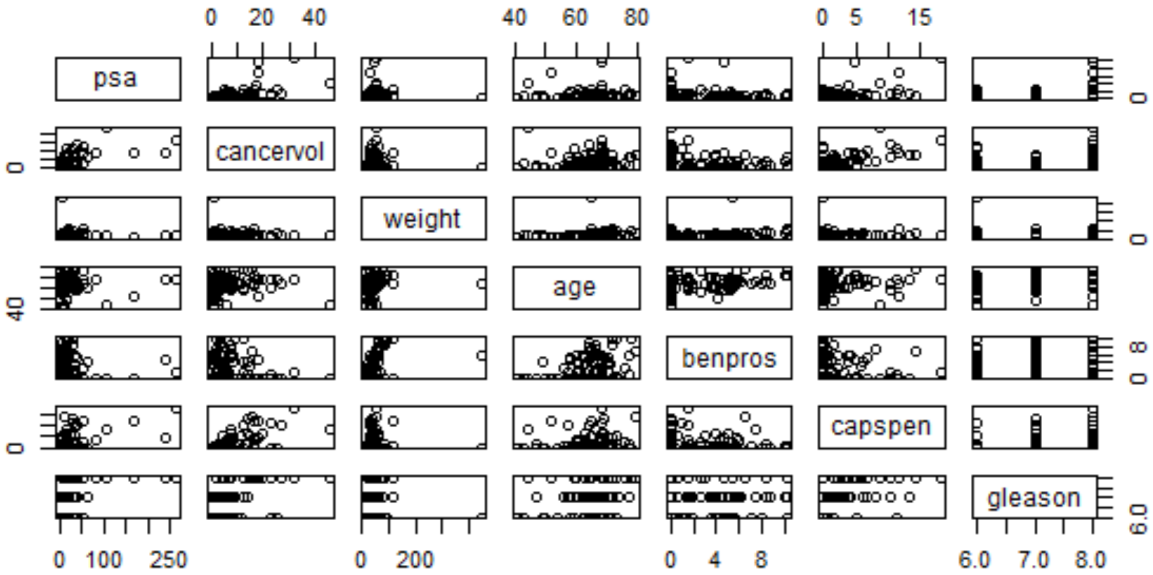

The histogram of capspen appears to be an exponential distribution which is very similar to the distribution of psa level, cancervol and benpros. Here, there might exist a strong linear relationship between PSA level and capspen.

7. Histogram of gleason:




As gleason is a quantitative variable, there are only 3 values (6,7,8) in the distribution as seen in the histogram.

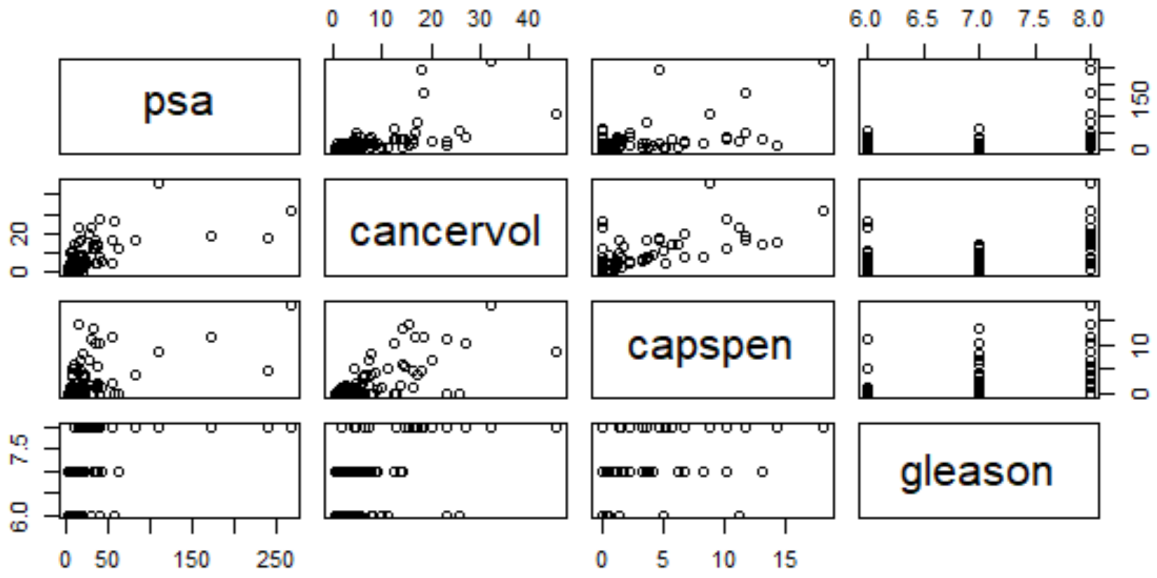
Before we start making our linear model, let us visualize the possible trends that exist from the scatterplots given below. The below scatterplots and correlations help us in better understanding the linear trends which may exist amongst the variables. Hence it becomes very essential to visualize these linear trends before we start building our linear model.




| | psa | cancervol | weight | age | benpros | vesinv | capspen | gleason |
|---|---|---|---|---|---|---|---|---|
| psa | 1.0000 | 0.6242 | 0.0262 | 0.0172 | -0.0165 | 0.5286 | 0.5508 | 0.4296 |
| cancervol | 0.6242 | 1.0000 | 0.0051 | 0.0391 | -0.1332 | 0.5817 | 0.6929 | 0.4814 |
| weight | 0.0262 | 0.0051 | 1.0000 | 0.1643 | 0.3218 | -0.0024 | 0.0016 | -0.0242 |
| age | 0.0172 | 0.0391 | 0.1643 | 1.0000 | 0.3663 | 0.1177 | 0.0996 | 0.2259 |
| benpros | -0.0165 | -0.1332 | 0.3218 | 0.3663 | 1.0000 | -0.1196 | -0.0830 | 0.0268 |
| vesinv | 0.5286 | 0.5817 | -0.0024 | 0.1177 | -0.1196 | 1.0000 | 0.6803 | 0.4286 |
| capspen | 0.5508 | 0.6929 | 0.0016 | 0.0996 | -0.0830 | 0.6803 | 1.0000 | 0.4616 |
| gleason | 0.4296 | 0.4814 | -0.0242 | 0.2259 | 0.0268 | 0.4286 | 0.4616 | 1.0000 |

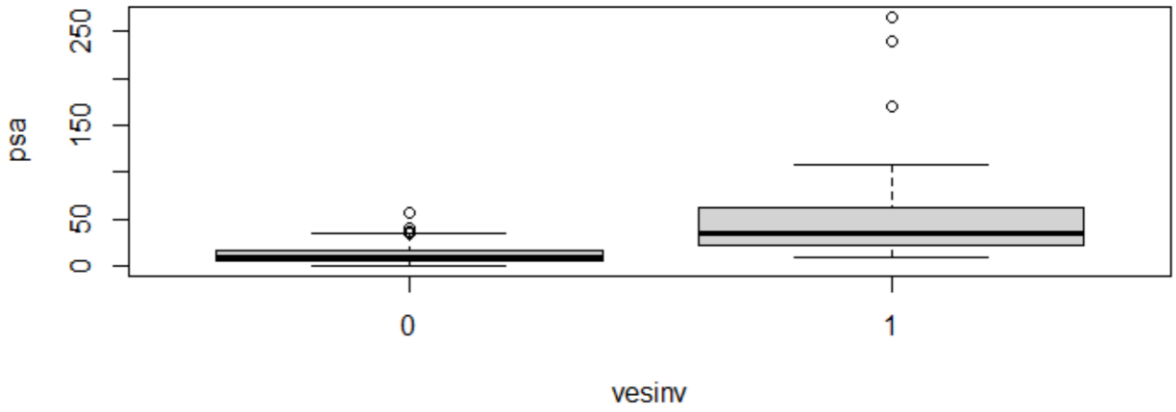
**Note: Here we ignore vesinv as it is a qualitative variable.**

The first row from the above result is very important, as it gives us the linear relationship of PSA levels with other variables. From the above data, we can conclude that there exists a negative correlation between the response variable PSA level and benpros. The linear relationship between the two is highly negative. However, the response variable PSA level shows a positive correlation and linear relationship with cancervol, capspen and gleason quantitative predictors. As, these predictors shows very high level of correlation with the response variable psa level, there are high chances of overfitting and which we should try avoiding. Let's observe a clearer vision of the linear trends between these variables:




We now perform an exploratory analysis of the qualitative variable.

Boxplot between psa level and qualitative variable vesinv:

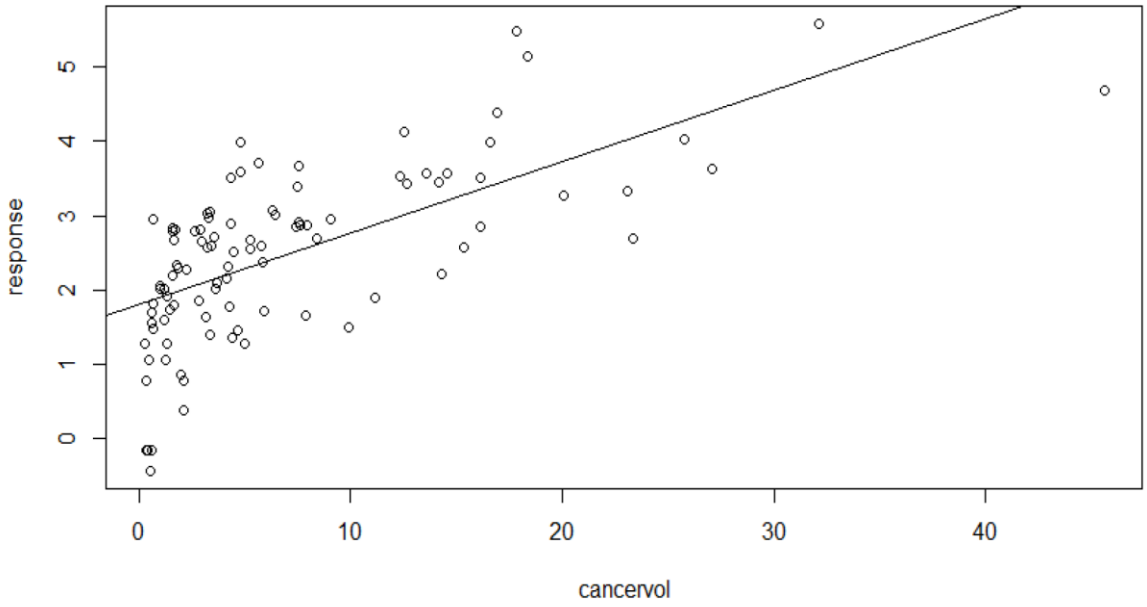



From the above boxplot we can conclude that psa values vary significantly over the levels of vesinv.

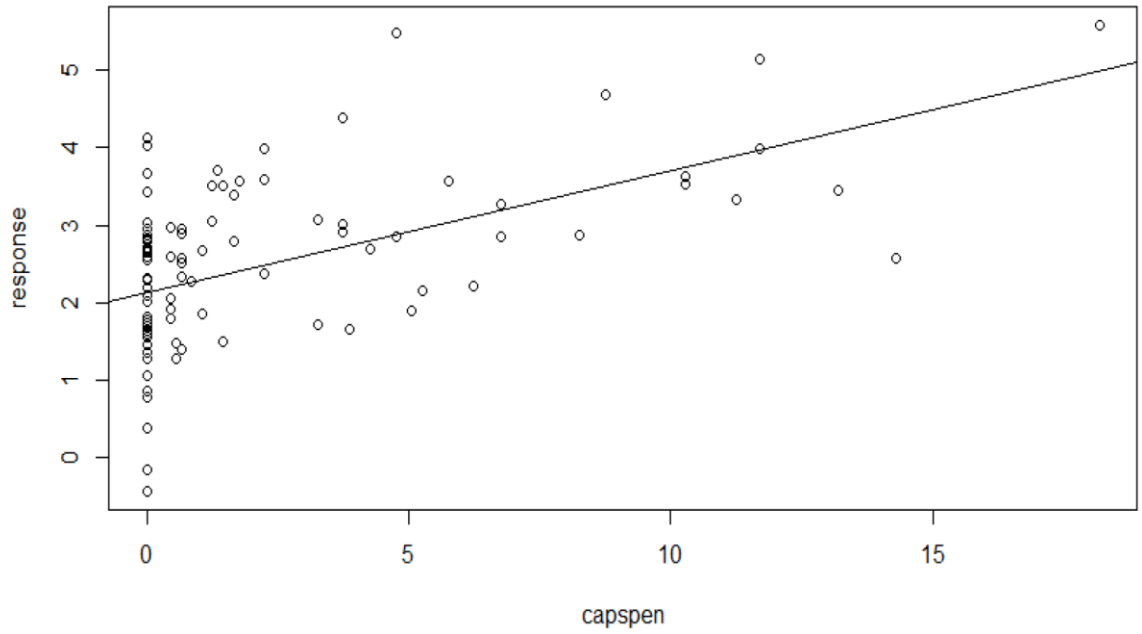
We have performed the exploratory analysis of both quantitative and qualitative variables. We can conclude that variables weight, age and benpros show no statistical evidence of including them into our preliminary linear model. Hence, we start building up our preliminary linear model using the quantitative variables such as cancervol, capspen and gleason and as discussed above that these predictors may cause an overfitting because of the strong correlation that exists between these predictors and the response variable. We transform the response to log(psa) by means of natural logarithmic transformation. So, let us observe the relationships between the transformed response log(psa) and each of these selected quantitative predictors.

The response along the y-axis shows the response variable(log(psa)):

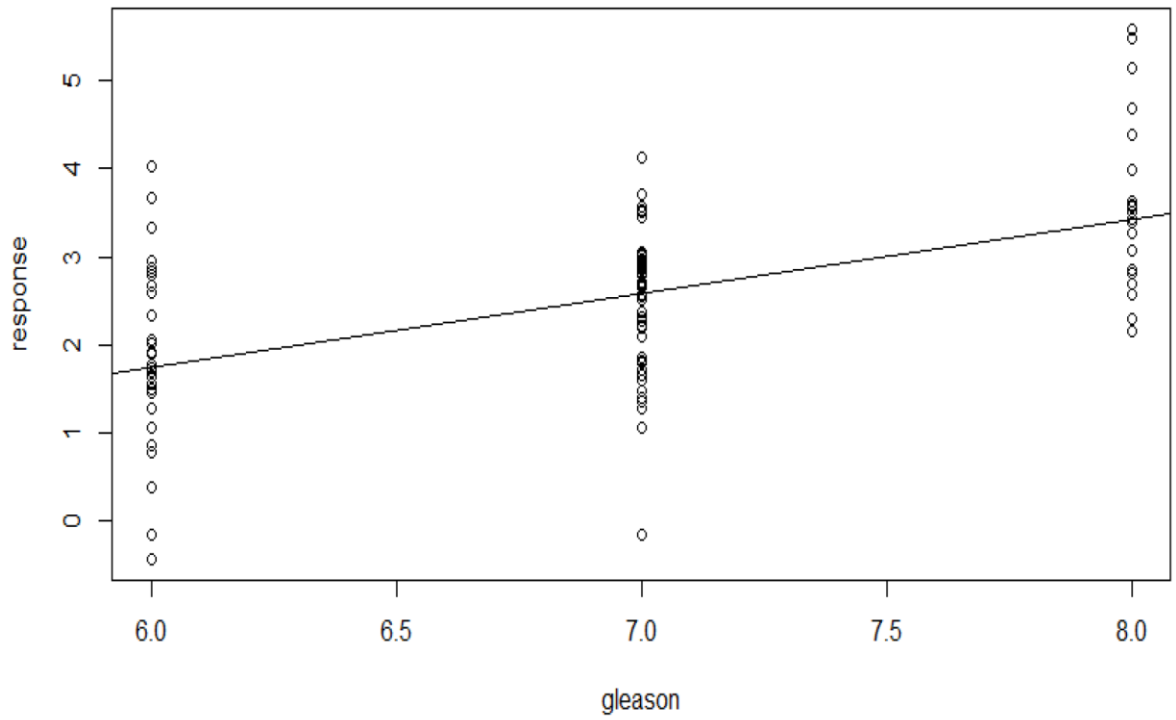
1. Response of cancervol:




2. Response of capspen:



3. Response of gleason:



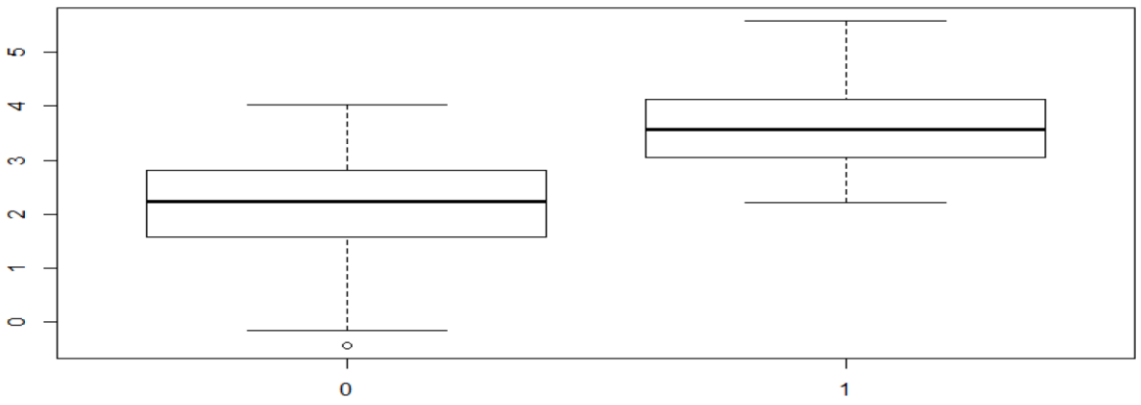

Let us review all the correlations between our transformed response variable log(psa) and other variables.

| | psa | cancervol | weight | age | benpros | capspen | gleason |
|---|---|---|---|---|---|---|---|
| psa | 1.0000 | 0.6571 | 0.1217 | 0.1699 | 0.1574 | 0.5180 | 0.5390 |

When we compare the correlations obtained from the previous results and the results from the data above, we observe a slight change, but our preliminary analysis of best possible predictors still holds as most part of it remains the same.

Qualitative predictor:

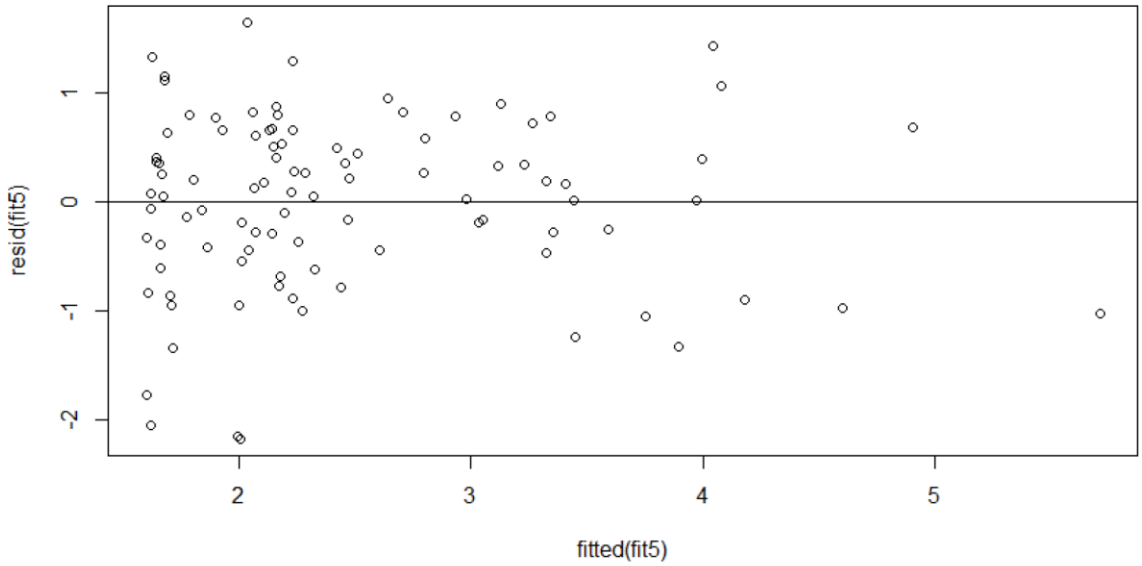
Boxplot of relationship between different levels of the qualitative variable (vesinv) and log(psa) is mapped again in order to ensure that the relationship still holds with newly transformed response.



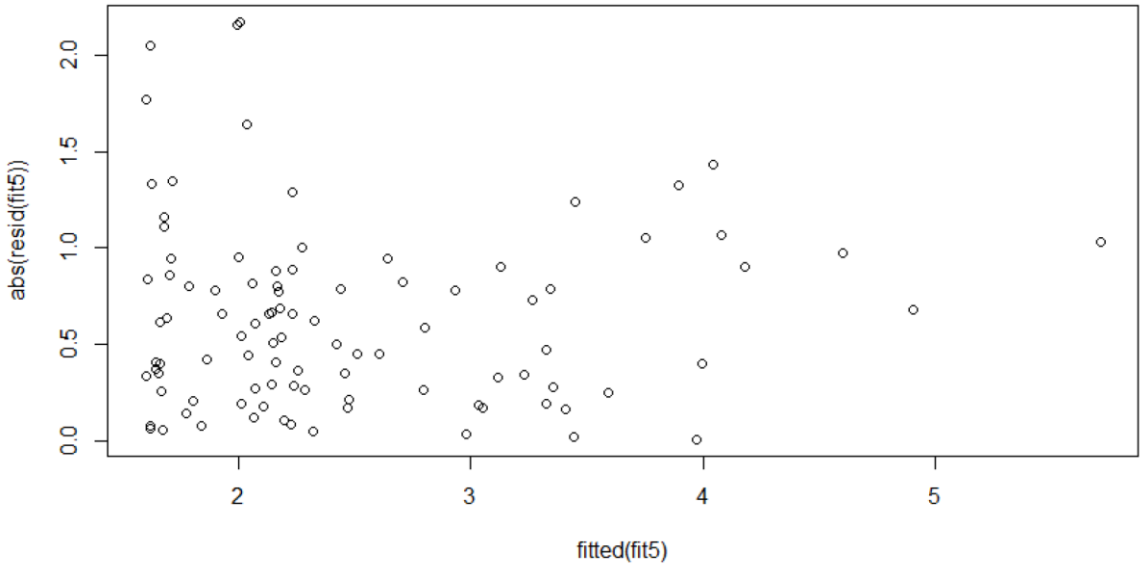

From the above boxplot we still observe some significant difference between the two levels even after the transformation.

As we observe a significant positive trend based on our exploratory analysis between the response variable log(psa), each variable and categorical predictor vesinv. We start building a linear model with quantitative predictors cancervol, capspen and gleason. We use all these predictors in our first model i.e. fit4. From the results of first model we get to know that the variable capspen is not needed and is redundant after observing the summary analysis and observing the t-test of the variable capspen. We also perform a partial F test for two nested models. If the p value is high for the partial F statistics, then we accept the null hypothesis or else we accept the alternative hypothesis. In the results obtained on performing F test for two nested models we obtain a high p value (0.4985), hence we accept the null hypothesis and based upon this we conclude that the reduced model is good. Thus, we can ignore the capspen variable and keep continuing our test on the other reduced model.
Now we further try to reduce our model by reducing the variable gleason and then the variable vesinv. On performing the summary analysis and partial F test we obtain a low p value for both these variables. Thus, we cannot ignore both these variable and we keep them in our model i.e. fit5, hence we use fit5 as our preliminary model.
On performing the automatic stepwise model selection procedures based on AIC and comparing our model (fit5) to them. We conclude that models produced by AIC methods- Forward Selection based on AIC (fit8_forward), Backward elimination based on AIC(fit9_backward) and the hybrid forward/backward(fit10_both) - are all the same with our preliminary model, with same quantitative predictors (cancervol and gleason) and the same categorical predictor(vesinv). We now perform a final summary analysis in order to prove that all the predictors are significant and finalise this model as our final model and perform diagnostics on it.

Residual Plot:

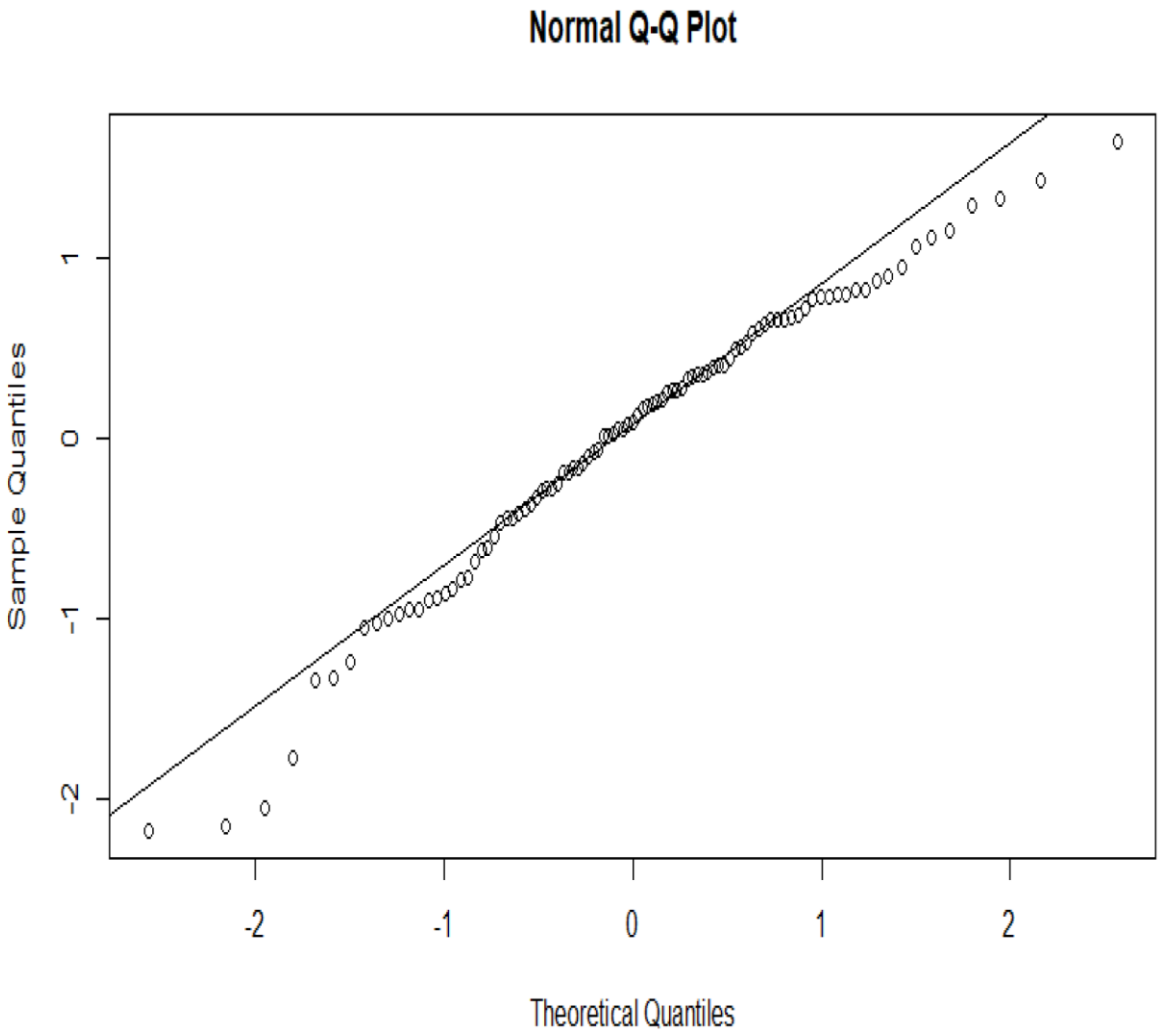



Absolute Residual Plot:




There is no particular trend observed in the residuals from the above plots.

Normal Q-Q Plot of Residuals:




From the Normal Q-Q plot of Residuals we can conclude that all of our model assumptions hold and hence the model which we have selected has passed all the diagnostics and we consider this model as our final model because the assumptions for Residuals also hold.
As all the assumptions for our final model holds true, we can use the results from our Model fit5 to predict the PSA level for patients whose quantitative predictors are at the sample means of the variable and qualitative predictors are at the most frequent category.

summary(fit5)

```
Call:
lm(formula = response ~ cancervol + gleason + vesinv)

Residuals:
     Min       1Q   Median       3Q      Max
-2.16928 -0.44558  0.08431  0.60719  1.64082

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72120    0.85749  -0.841   0.4025
cancervol    0.05981    0.01352   4.425 2.62e-05 ***
gleason      0.38491    0.12966   2.969   0.0038 **
vesinv       0.62117    0.24962   2.488   0.0146 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared:  0.5277,   Adjusted R-squared:  0.5125
F-statistic: 34.64 on 3 and 93 DF,  p-value: 4.022e-15
```

From the equation given below and the results obtained from the fit5 Model, we predict the PSA Level.

Y = β0 + β1 * cancervol + β2 * gleason + β3 * vesinv

Hence, the predicted PSA level for patients whose quantitative predictors are at the sample means of the variable and qualitative predictors are at the most frequent category is 2.34414346682474.

R CODE:

```
> # Reading the prostate_cancer.csv file
> CollectionData = read.csv('C:/Users/Nikita/Desktop/New folder/Personal Information/Statistical Methods for Data Science/Mini Project 6/prostate_cancer.csv')
> psa = CollectionData[,2]
> cancervol = CollectionData[,3]
> weight = CollectionData[,4]
> age = CollectionData[,5]
> benpros = CollectionData[,6]
> vesinv = CollectionData[,7]
> capspen = CollectionData[,8]
> gleason = CollectionData[,9]
> # Performing Exploratory Analysis of Response(i.e PSA)
> # 1. Histogram
> hist(psa,xlab="PSA Level",main="HISTOGRAM OF PSA LEVEL",col="yellow", breaks=20)
> # 2. Boxplot
> boxplot(psa)
> # 3. Boxplot of transformed response i.e natural log transformation of PSA
> boxplot(log(psa))
> # 4. Q-Q Plots
> qqnorm(psa)
> qqline(psa)
> # 1. Histogram of each variable using single for loop
> for (i in 1:9) {
+ hist(CollectionData[,i], xlab=colnames(CollectionData)[i],
+ main=paste("Histogram of",colnames(CollectionData[i])),
+ col="red",breaks=20)
+ }
> # 2. Scatterplots and Correlation between PSA and other variables
> # 2.a Without any Transformation of PSA Level
> pairs(~psa + cancervol + weight + age + benpros + capspen + gleason, data = CollectionData)
> # 2.b With Natural log Transformed of PSA Level
> pairs(~psa + cancervol + capspen + gleason, data = CollectionData)
> # Obtaining all the correlations between each pair of variables
> prostatecancer_cor = cor(CollectionData[,2:9])
> round(prostatecancer_cor,4)
psa cancervol weight age benpros vesinv capspen gleason
psa 1.0000 0.6242 0.0262 0.0172 -0.0165 0.5286 0.5508 0.4296
cancervol 0.6242 1.0000 0.0051 0.0391 -0.1332 0.5817 0.6929 0.4814
weight 0.0262 0.0051 1.0000 0.1643 0.3218 -0.0024 0.0016 -0.0242
age 0.0172 0.0391 0.1643 1.0000 0.3663 0.1177 0.0996 0.2259
benpros -0.0165 -0.1332 0.3218 0.3663 1.0000 -0.1196 -0.0830 0.0268
vesinv 0.5286 0.5817 -0.0024 0.1177 -0.1196 1.0000 0.6803 0.4286
capspen 0.5508 0.6929 0.0016 0.0996 -0.0830 0.6803 1.0000 0.4616
gleason 0.4296 0.4814 -0.0242 0.2259 0.0268 0.4286 0.4616 1.0000
> # Obtaining correlation between Natural log Transofrmation of PSA and other vairables
> cor(CollectionData,log(psa))
[,1]
```

```
subject 0.9581180
psa 0.7210677
cancervol 0.6570739
weight 0.1217208

> # Boxplots
> boxplot(psa~vesinv)
> # As Vesinv is a qualitative data we take vesinv as a categorical variable
> CollectionData$vesinv = factor(CollectionData$vesinv)
> response = log(psa)
> # cancervol and response
> plot(cancervol,response)
> fit1 = lm(response~cancervol,data=CollectionData)
> abline(fit1)
> # capspen and response
> plot(capspen,response)
> fit2 = lm(response~capspen,data=CollectionData)
> abline(fit2)
> # gleason and response
> plot(gleason,response)
> fit3 = lm(response~gleason,data=CollectionData)
> abline(fit3)
> # Checking correlation between newly transformed response and other variables
> NewData = CollectionData
> CollectionData$psa = log(psa)
> prostatecancer_cor = cor(CollectionData[c(2,3,4,5,6,8,9)])
> round(prostatecancer_cor,4)
psa cancervol weight age benpros capspen gleason
psa 1.0000 0.6571 0.1217 0.1699 0.1574 0.5180 0.5390
cancervol 0.6571 1.0000 0.0051 0.0391 -0.1332 0.6929 0.4814
weight 0.1217 0.0051 1.0000 0.1643 0.3218 0.0016 -0.0242
age 0.1699 0.0391 0.1643 1.0000 0.3663 0.0996 0.2259
benpros 0.1574 -0.1332 0.3218 0.3663 1.0000 -0.0830 0.0268
capspen 0.5180 0.6929 0.0016 0.0996 -0.0830 1.0000 0.4616
gleason 0.5390 0.4814 -0.0242 0.2259 0.0268 0.4616 1.0000
> # Qualitative Exploratory Analysis of Newly Transformed Response
> #Boxplot
> boxplot(response~vesinv)
> # Creating first with quantitative variables and qualitative variables
> # Obtaining response when we use all three variables
> fit4 = lm(response~cancervol+capspen+gleason+vesinv)
> fit4
Call:
lm(formula = response ~ cancervol + capspen + gleason + vesinv)
Coefficients:
(Intercept) cancervol capspen gleason vesinv
-0.79386 0.06452 -0.02348 0.39566 0.70675
> # Summary for the model
> summary(fit4)

Call:
```

```
lm(formula = response ~ cancervol + capspen + gleason + vesinv)
Residuals:
Min 1Q Median 3Q Max
-2.1747 -0.4497 0.1049 0.6215 1.6135
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.79386 0.86660 -0.916 0.36203
cancervol 0.06452 0.01522 4.238 5.35e-05 ***
capspen -0.02348 0.03455 -0.680 0.49852
gleason 0.39566 0.13100 3.020 0.00327 **
vesinv 0.70675 0.28024 2.522 0.01339 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8078 on 92 degrees of freedom
Multiple R-squared: 0.5301, Adjusted R-squared: 0.5097
F-statistic: 25.95 on 4 and 92 DF, p-value: 2.075e-14
> # Removing capspen from the model
> fit5 = lm(response~cancervol + gleason + vesinv)
> # Removing capsen and gleason
> fit6 = lm(response~cancervol + vesinv)
> # Performing partial F test to check the significance of capspen (fit4, fit5)
> anova(fit4,fit5)
Analysis of Variance Table
Model 1: response ~ cancervol + capspen + gleason + vesinv
Model 2: response ~ cancervol + gleason + vesinv
Res.Df RSS Df Sum of Sq F Pr(>F)
1 92 60.039
2 93 60.340 -1 -0.30134 0.4617 0.4985
> # Performing partial F test to check the significance of capspen (fit5, fit6) if gleason is required
> anova(fit5,fit6)
Analysis of Variance Table
Model 1: response ~ cancervol + gleason + vesinv
Model 2: response ~ cancervol + vesinv
Res.Df RSS Df Sum of Sq F Pr(>F)
1 93 60.340
2 94 66.058 -1 -5.7179 8.8127 0.003804 **
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> # Validating if categorical variable vesinv is an important predictor or not and performing partial F
test to check its significance
> fit7 = lm(response~cancervol + gleason)
> anova(fit6,fit7)
Analysis of Variance Table
Model 1: response ~ cancervol + vesinv
Model 2: response ~ cancervol + gleason
```

```
Res.Df RSS Df Sum of Sq F Pr(>F)
1 94 66.058
2 94 64.358 0 1.7
> # Based on the above results we choose fit5 as preliminary model
> summary(fit5)
Call:
lm(formula = response ~ cancervol + gleason + vesinv)
Residuals:
Min 1Q Median 3Q Max
-2.16928 -0.44558 0.08431 0.60719 1.64082
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72120 0.85749 -0.841 0.4025
cancervol 0.05981 0.01352 4.425 2.62e-05 ***
gleason 0.38491 0.12966 2.969 0.0038 **
vesinv 0.62117 0.24962 2.488 0.0146 *

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared: 0.5277, Adjusted R-squared: 0.5125
F-statistic: 34.64 on 3 and 93 DF, p-value: 4.022e-15
> fit8_forward = step(lm(response ~ 1, data = NewData), scope = list(upper =
~cancervol+capspen+gleason+vesinv),
+ direction = "forward")
Start: AIC=28.72
response ~ 1
Df Sum of Sq RSS AIC
+ cancervol 1 55.164 72.605 -24.0986
+ vesinv 1 40.984 86.785 -6.7944
+ gleason 1 37.122 90.647 -2.5707
+ capspen 1 34.286 93.482 0.4169
<none> 127.769 28.7246
Step: AIC=-24.1
response ~ cancervol
Df Sum of Sq RSS AIC
+ gleason 1 8.2468 64.358 -33.794
+ vesinv 1 6.5468 66.058 -31.265
<none> 72.605 -24.099
+ capspen 1 0.9673 71.638 -23.400
Step: AIC=-33.79
response ~ cancervol + gleason
Df Sum of Sq RSS AIC
+ vesinv 1 4.0178 60.340 -38.047
```

```
<none> 64.358 -33.794
+ capspen 1 0.1685 64.190 -32.048
Step: AIC=-38.05
response ~ cancervol + gleason + vesinv
Df Sum of Sq RSS AIC
<none> 60.340 -38.047
+ capspen 1 0.30134 60.039 -36.532
> fit9_backward = step(lm(response~cancervol+capspen+gleason+vesinv, data = NewData), scope =
list(lower = ~1),
+ direction = "backward")
Start: AIC=-36.53
response ~ cancervol + capspen + gleason + vesinv
Df Sum of Sq RSS AIC
- capspen 1 0.3013 60.340 -38.047
<none> 60.039 -36.532
- vesinv 1 4.1507 64.190 -32.048
- gleason 1 5.9535 65.993 -29.361
- cancervol 1 11.7209 71.760 -21.234
Step: AIC=-38.05
response ~ cancervol + gleason + vesinv
Df Sum of Sq RSS AIC
<none> 60.340 -38.047
- vesinv 1 4.0178 64.358 -33.794
- gleason 1 5.7179 66.058 -31.265
- cancervol 1 12.7041 73.044 -21.513
> fit10_both = step(lm(response ~ 1, data = NewData), scope = list(lower = ~1, upper =
~cancervol+capspen+gleason+vesinv),
+ direction = "both")
Start: AIC=28.72
response ~ 1
Df Sum of Sq RSS AIC
+ cancervol 1 55.164 72.605 -24.0986
+ vesinv 1 40.984 86.785 -6.7944
+ gleason 1 37.122 90.647 -2.5707
+ capspen 1 34.286 93.482 0.4169
<none> 127.769 28.7246
Step: AIC=-24.1
response ~ cancervol
Df Sum of Sq RSS AIC
+ gleason 1 8.247 64.358 -33.794
+ vesinv 1 6.547 66.058 -31.265
<none> 72.605 -24.099
+ capspen 1 0.967 71.638 -23.400
- cancervol 1 55.164 127.769 28.725
```

```
Step: AIC=-33.79
response ~ cancervol + gleason
Df Sum of Sq RSS AIC
+ vesinv 1 4.0178 60.340 -38.047
<none> 64.358 -33.794
+ capspen 1 0.1685 64.190 -32.048
- gleason 1 8.2468 72.605 -24.099
- cancervol 1 26.2887 90.647 -2.571
Step: AIC=-38.05
response ~ cancervol + gleason + vesinv
Df Sum of Sq RSS AIC
<none> 60.340 -38.047
+ capspen 1 0.3013 60.039 -36.532
- vesinv 1 4.0178 64.358 -33.794
- gleason 1 5.7179 66.058 -31.265
- cancervol 1 12.7041 73.044 -21.513
> summary(fit5)
Call:
lm(formula = response ~ cancervol + gleason + vesinv)
Residuals:
Min 1Q Median 3Q Max
-2.16928 -0.44558 0.08431 0.60719 1.64082
Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.72120 0.85749 -0.841 0.4025
cancervol 0.05981 0.01352 4.425 2.62e-05 ***
gleason 0.38491 0.12966 2.969 0.0038 **
vesinv 0.62117 0.24962 2.488 0.0146 *
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.8055 on 93 degrees of freedom
Multiple R-squared: 0.5277, Adjusted R-squared: 0.5125
F-statistic: 34.64 on 3 and 93 DF, p-value: 4.022e-15
> # Residual plot
> plot(fitted(fit5), resid(fit5))
> abline(h = 0)
> # Absolute residuals plot
> plot(fitted(fit5), abs(resid(fit5)))
> # Normal QQ plot
> qqnorm(resid(fit5))
> qqline(resid(fit5))
>
> #All assumptions hold
> # This preliminary model passes the diagnostics. So, we take this as our final model.
> # Now we ouse our final model i.e fit 5 to predict the PSA level for a patient
```

```
> # We know that the equation is y = Intercept (Beta0) + Beta 1 * cancervol + Beta2 * gleason + Beta3
* vesinv
> cancervol_mean = mean(cancervol)
> gleason_mean = mean(gleason)
> predicted_PSALevel = -0.72120 + 0.05981 * cancervol_mean + 0.38491 * gleason_mean + 0.62117
* 0
```